Figure 1.

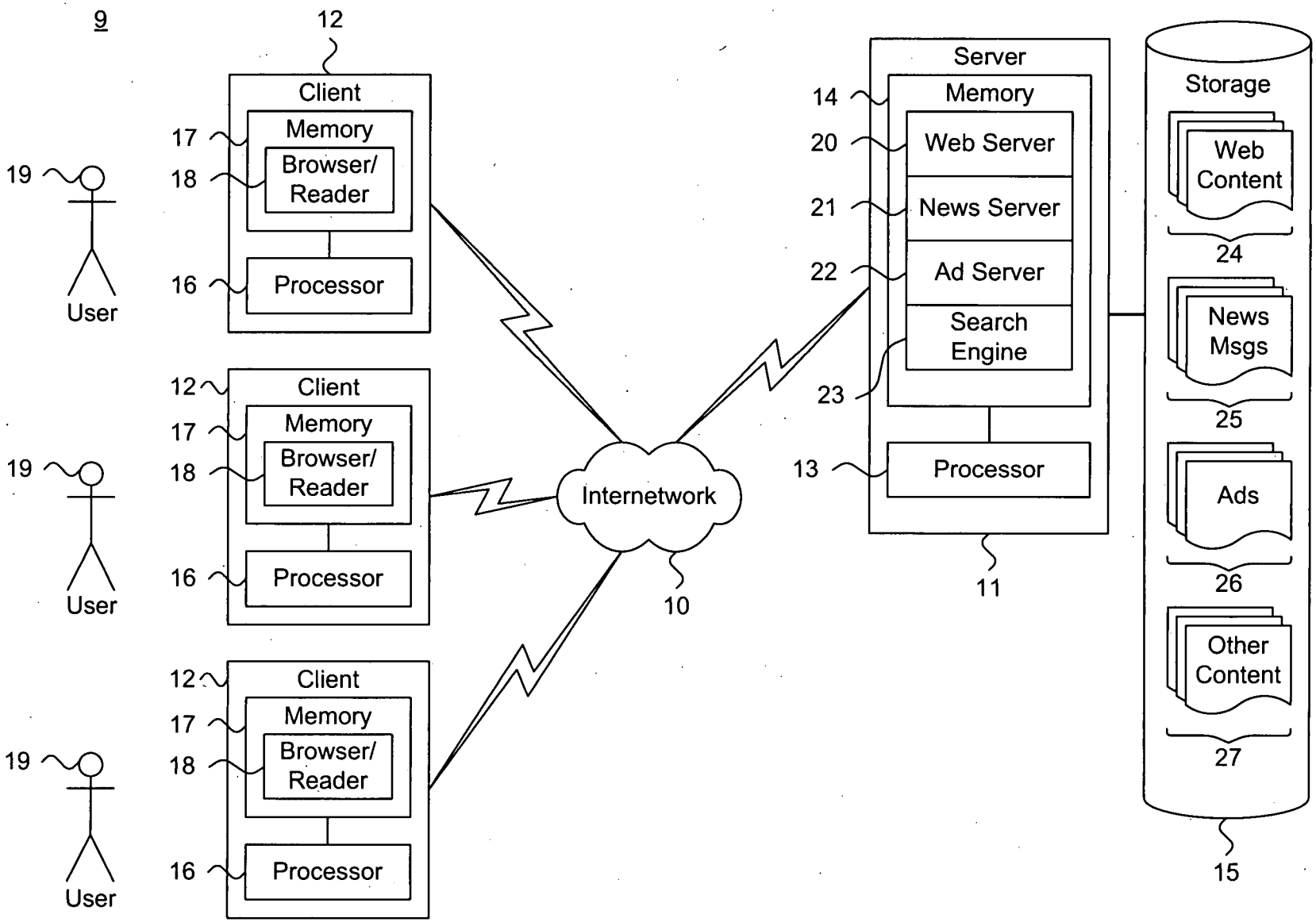
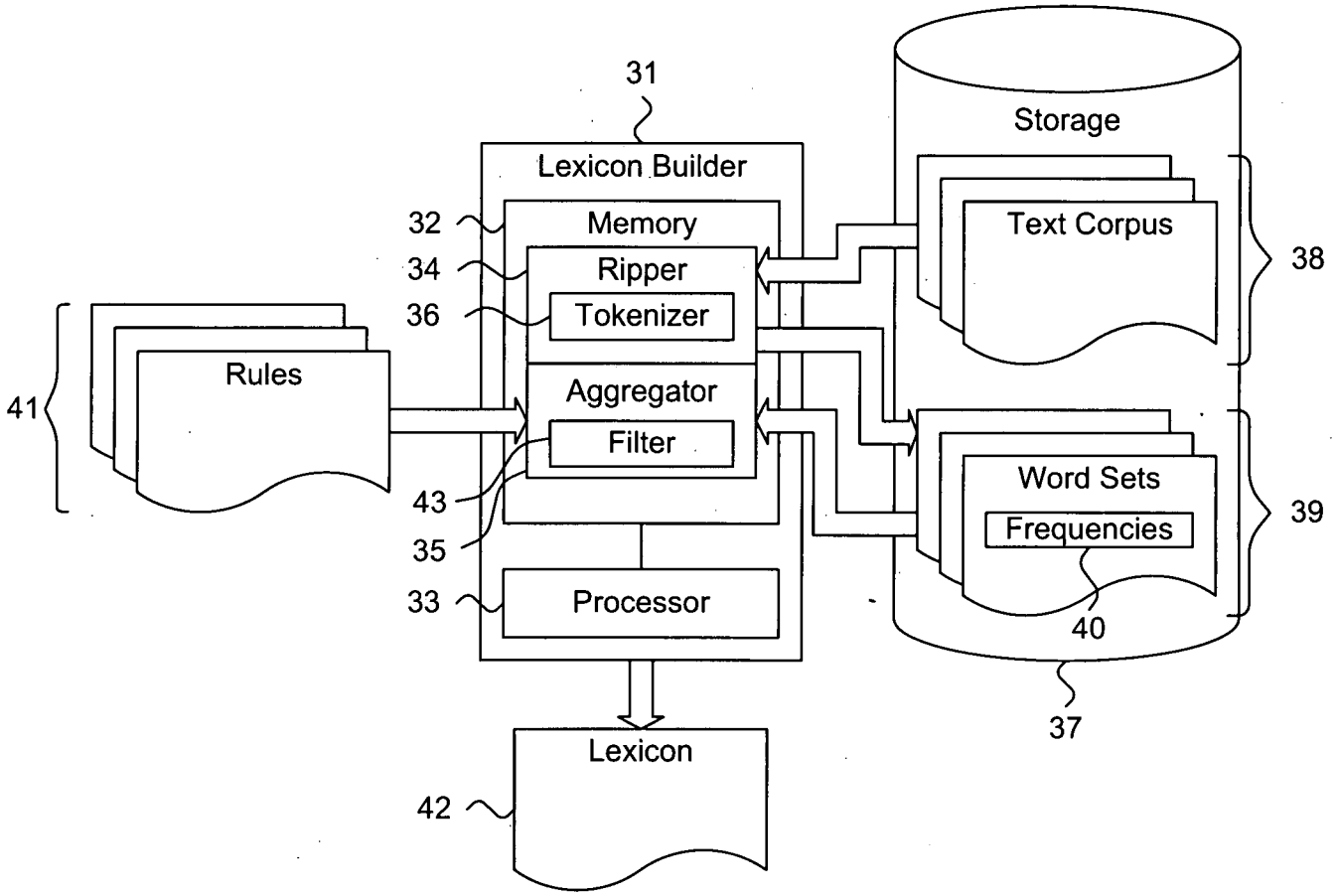

9
12
Client
Memory
Browser/
Reader
Processor
17
18
16
19
User
Internetwork
10
Server
Memory
Web Server
News Server
Ad Server
Search
Engine
Processor
14
20
21
22
23
13
11
Storage
Web
Content
24
News
Msgs
25
Ads
26
Other
Content
27
15

Figure 2.


30
31
Lexicon Builder
32
Memory
34
Ripper
36
Tokenizer
Aggregator
Filter
43
35
33
Processor
41
Rules
Storage
Text Corpus
38
Word Sets
Frequencies
39
40
37
Lexicon
42

Figur 3.

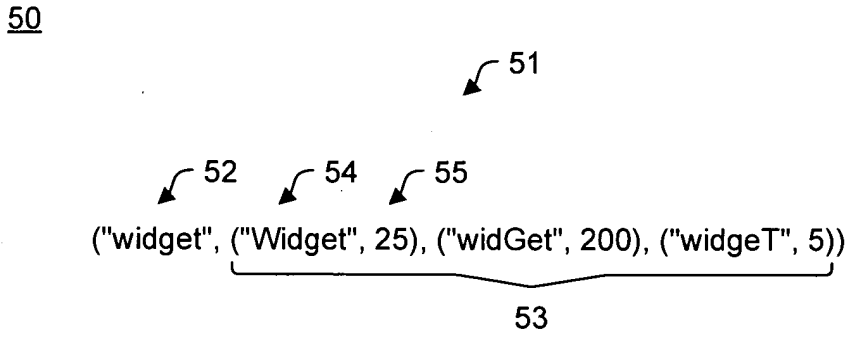


Figure 5.

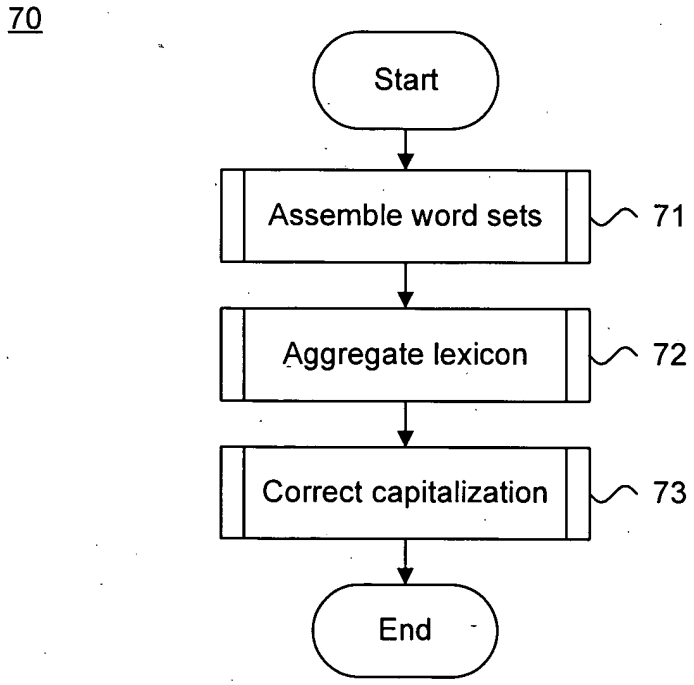


Figure 4.

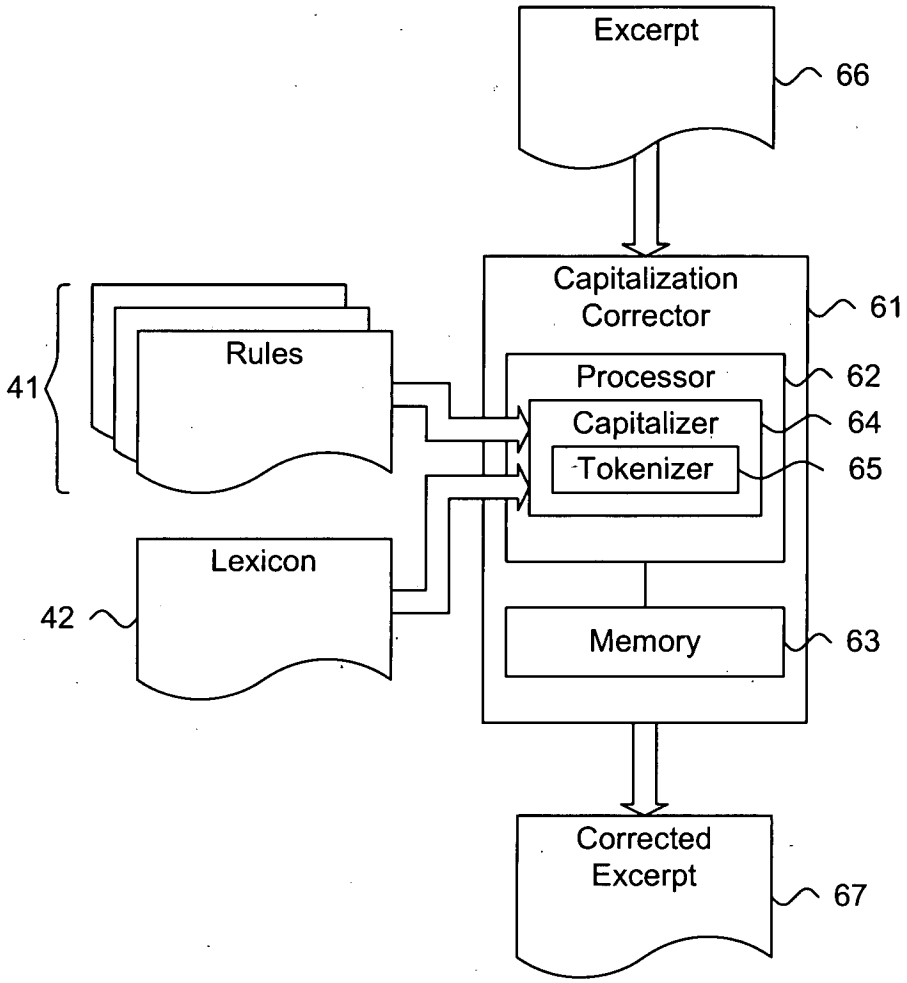

60
Excerpt
66
Capitalization Corrector
61
Rules
41
Processor
62
Capitalizer
64
Tokenizer
65
Lexicon
42
Memory
63
Corrected Excerpt
67

Figure 6.

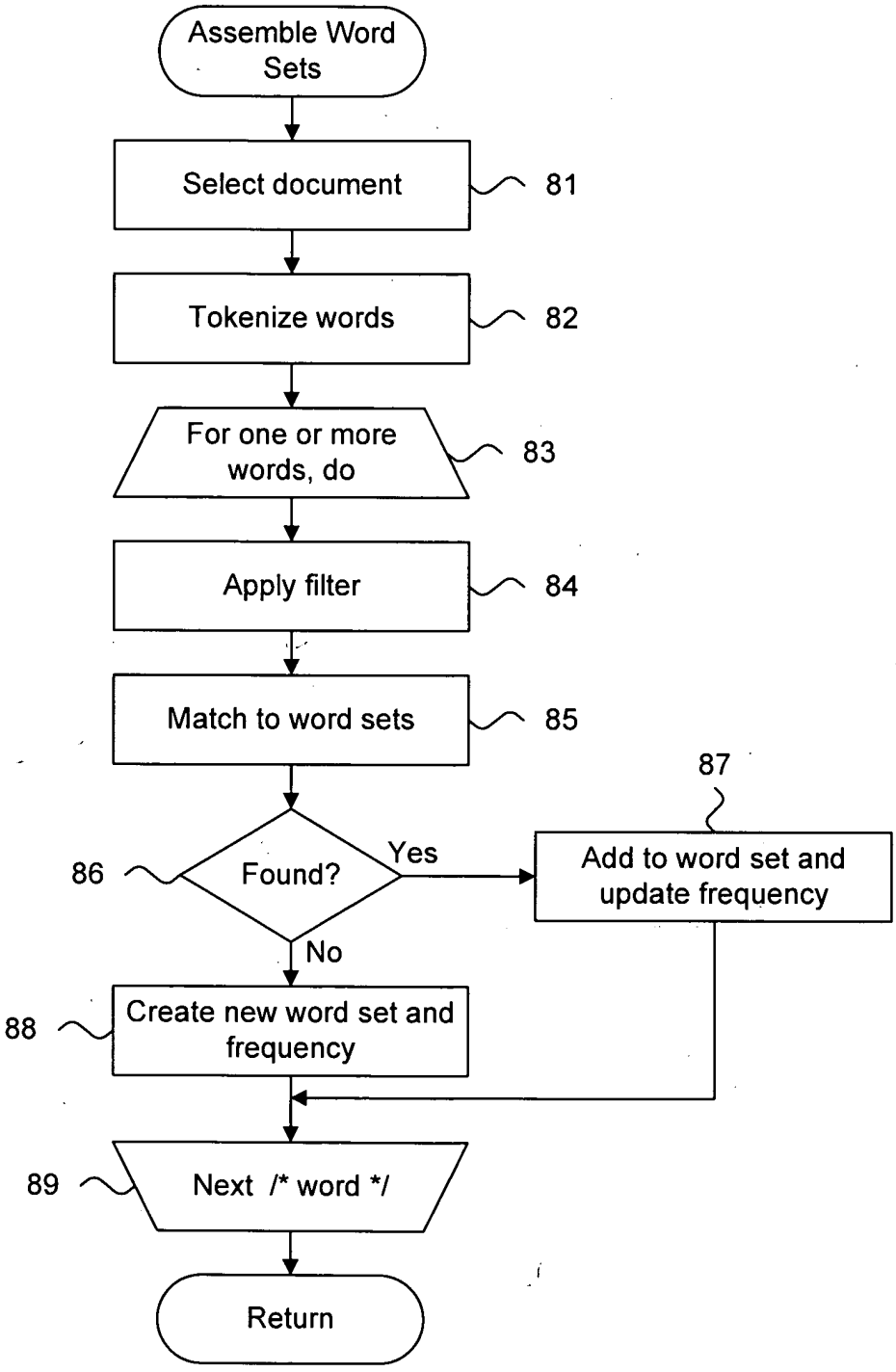
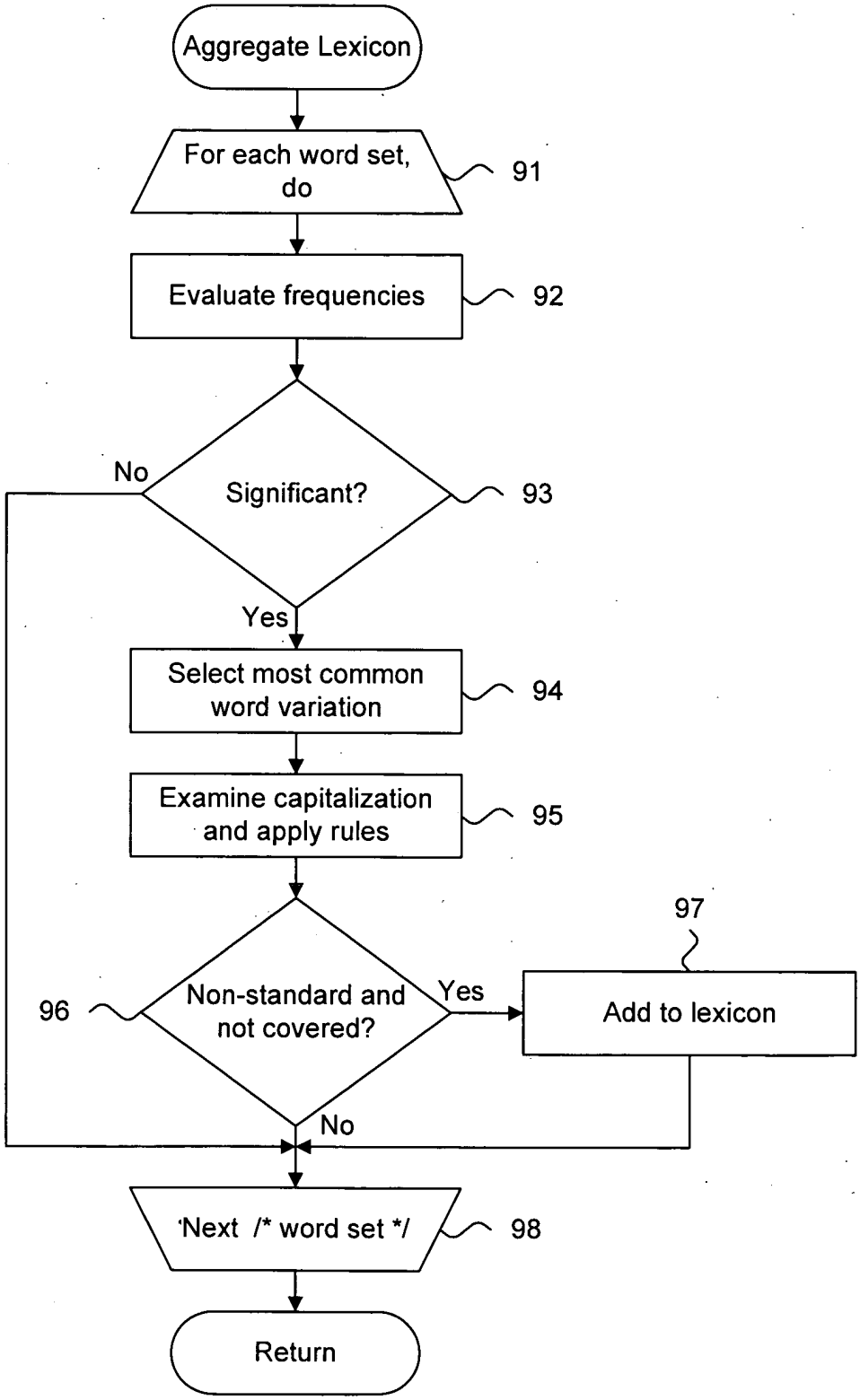
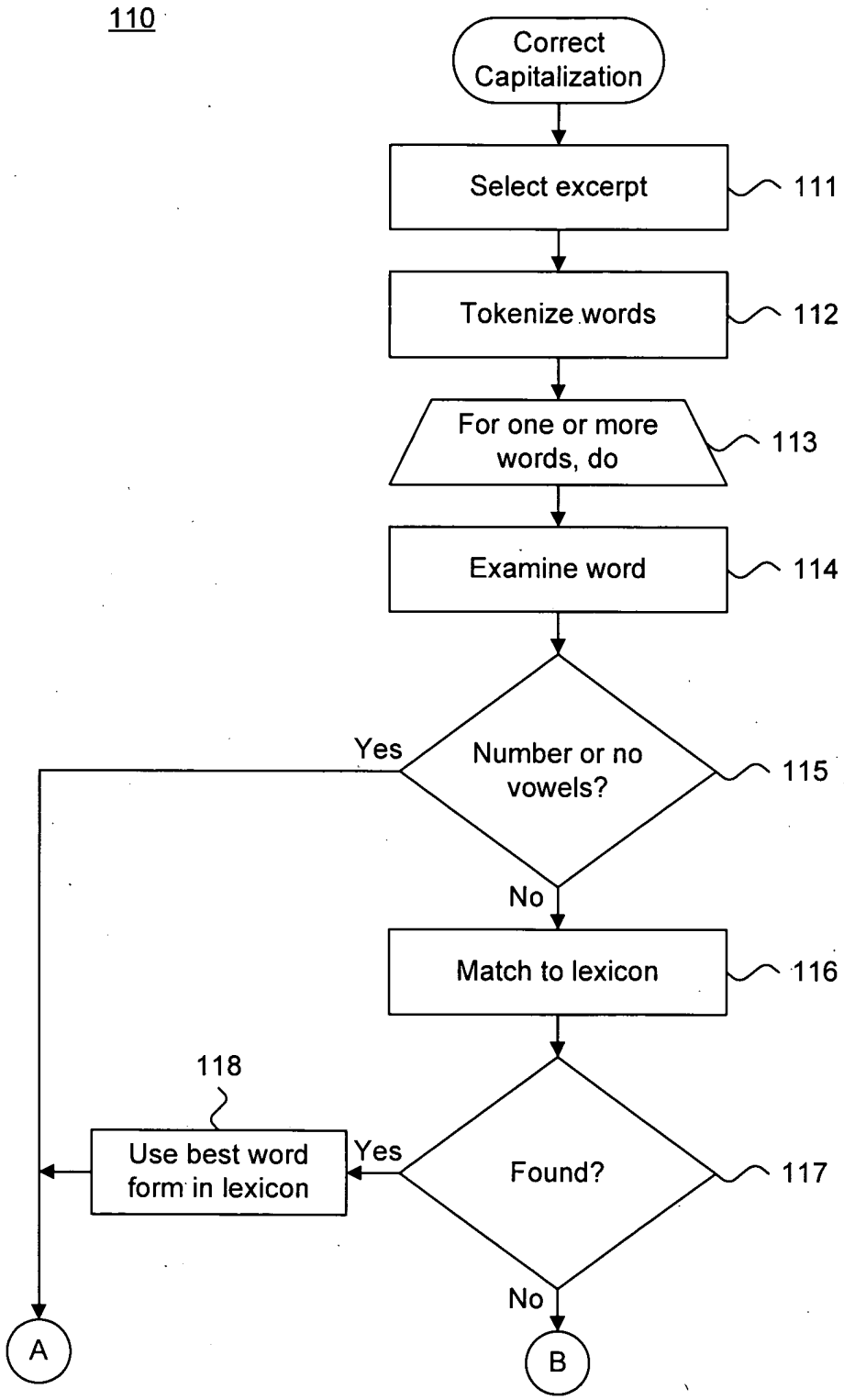

80
Assemble Word Sets
Select document
81
Tokenize words
82
For one or more words, do
83
Apply filter
84
Match to word sets
85
87
86
Found?
Yes
Add to word set and update frequency
No
88
Create new word set and frequency
89
Next /* word */
Return

Figure 7.


90
Aggregate Lexicon
For each word set, do
91
Evaluate frequencies
92
No
Significant?
93
Yes
Select most common word variation
94
Examine capitalization and apply rules
95
97
96
Non-standard and not covered?
Yes
Add to lexicon
No
'Next /* word set */
98
Return

Figur 8.


110
Correct Capitalization
Select excerpt
111
Tokenize words
112
For one or more words, do
113
Examine word
114
Number or no vowels?
115
Yes
No
Match to lexicon
116
Found?
117
Yes
No
118
Use best word form in lexicon
A
B

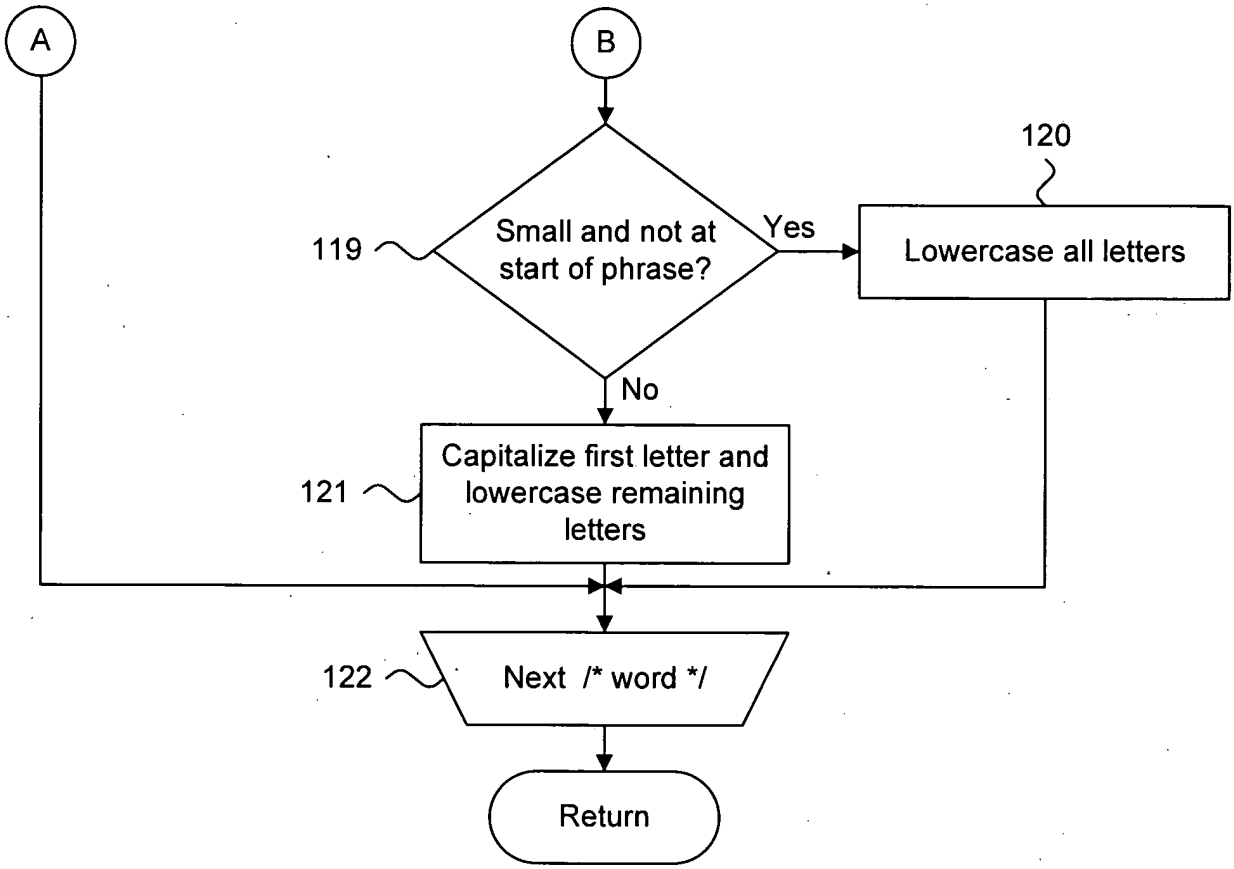



Figure 8 (Cont).